

Weekly Report

May 19, 2019

1 Work

1. 本周在进行unpair setting下的图片增强，本周进行了我们的第二阶段，目前效果仍不理想，但是发现了需要改进的方向。
2. Adversarial Attack使用字典直接学习对抗样本，使用了两种优化方式：1) ADAM等梯度下降法；2) L-BFGS，即每个样本都会搜索到局部最优解。目前发现方法1效率更好一些，方法二特别缓慢，随着优化到前一个样本的最优，后一个样本就会需要更多时间优化。之后将测试方法1的各种性能。目前的进度来说，难以赶上NIPS，可能再往后面的投稿。
3. 工作时长：工作日每天9个小时，周末共10个小时，共55个小时。

1.1 工作进度

Table 1: 工作进度

项目	进度	截止时间
DRGraph	正在修改代码	6.30
unpair 低光照图片增强	目前初步的实验效果不佳	7.30
Universal Flow Attack	基于字典学习Adversarial Attack	6.30

2 Paper Reading

2.1 PLAYING THE GAME OF UNIVERSAL ADVERSARIAL PERTURBATIONS

本文讲对抗样本的生成和防御看做是一个对抗游戏，生成对抗样本就是找到一个对于目前所有的分类器攻击效果最好的对抗样本，防御就是找到一个新的分类器，它可以防御更多的对抗攻击。

2.2 Adversarial Transformation Networks: Learning to Generate Adversarial Examples

使用神经网络生成对抗样本

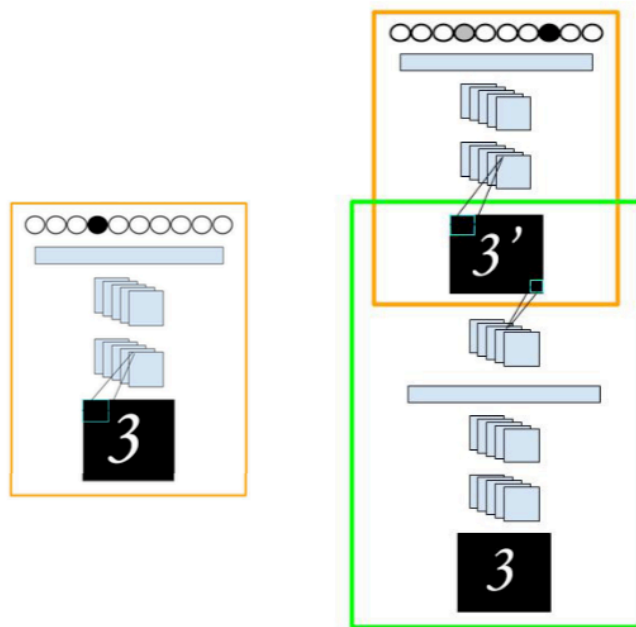


Figure 1: #2

2.3 N ATTACK: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks

对于针对black box的模型的攻击（基于梯度），因为不能得知内部的函数情况，所以估计的梯度会有不稳定等现象。本文提出一个平滑的目标函数，因此使得对抗样本可以在输入数据的一个局部分布中采样得到。因此原始函数梯度转而变成分布的参数的梯度，更容易求得。

2.4 Leveraging Large-Scale Uncurated Data for Unsupervised Pre-training of Visual Features

通过无监督的方法抽取特征可以用于多种任务，然而当前方法在一些没有经过人工筛选的数据集上的性能不佳。本文提出了一个基于Self-supervision 和 Deep clustering 方法的无监督学习方法，即创造一些label用于监督学习（Self-supervision 用旋转角度作为label，而Deep clustering 使用kmean聚类作为label，两个label的笛卡尔乘积作为新的label）。

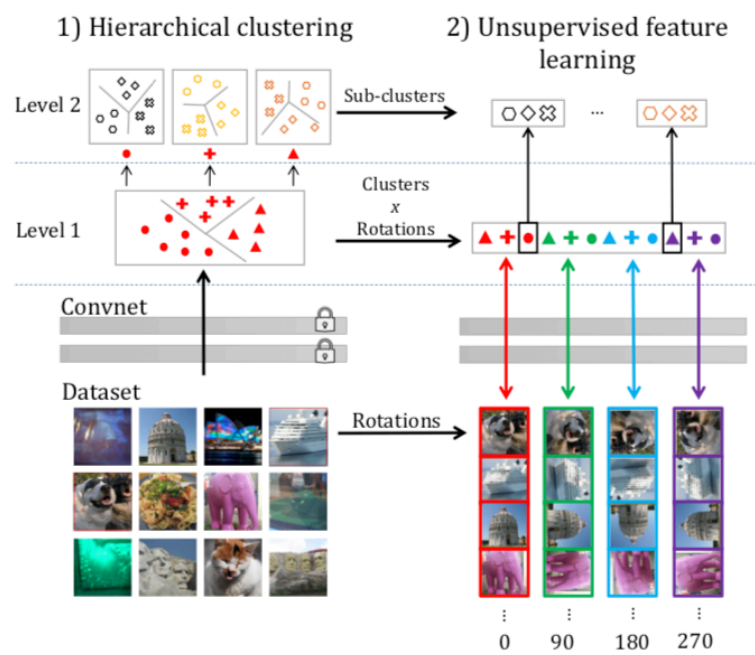


Figure 2: #4